PREDICTING DEVICE PERFORMING PREDICTION BASED ON ANALOGOUS EXAMPLE AND METHOD

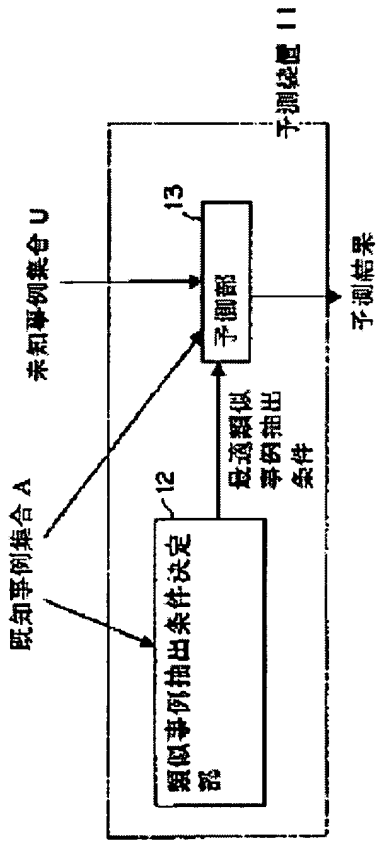
Patent number: JP2000155681
Publication date: 2000-06-06
Inventor: MAEDA KAZUO; YAGINUMA YOSHINORI
Applicant: FUJITSU LTD
Classification:
- international: ***G06N5/00;*** **G06N5/00**; (IPC1-7): G06F9/44; G06F9/44
- european: G06N5/00
Application number: JP19980332503 19981124
Priority number(s): JP19980332503 19981124

Also published as:
US 6418425 (B1)

Report a data error here

Abstract of **JP2000155681**

PROBLEM TO BE SOLVED: To perform prediction based on an analogous example fast and also with high accuracy. SOLUTION: An analogous example extraction condition deciding part 12 automatically decides an optimum analogous example extraction condition by using a known example set A, and a predicting part 13 predicts an unknown field of an unknown example set U from the known example set A by using the condition. In such a case, the predicting part 13 calculates the similarity depending on the distribution of unknown field values in the set A and extracts an analogous example set based on the similarity. Also, the predicting part 13 stops a similarity calculation at the point of time when it is determined that the similarity does not meet a prescribed condition.



(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号
特開2000－155681
（P2000－155681A）
(43)公開日　平成12年６月６日(2000.6.6)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｇ０６Ｆ　9/44	５５０	Ｇ０６Ｆ　9/44	５５０Ｎ	
	５８０		５８０Ｊ	

審査請求　未請求　請求項の数33　ＯＬ　（全 26 頁）

(21)出願番号　特願平10－332503

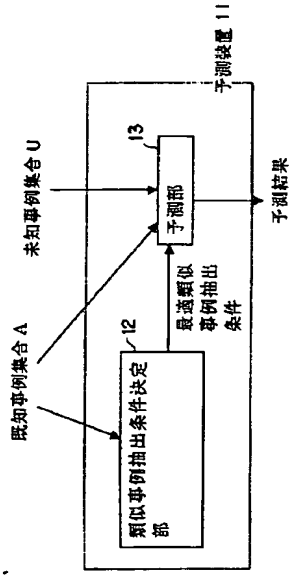
(22)出願日　平成10年11月24日(1998.11.24)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　前田　一穂
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　柳沼　義典
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100074099
弁理士　大菅　義之　（外１名）

(54)【発明の名称】　類似事例に基づく予測を行う予測装置および方法

(57)【要約】
【課題】　類似事例に基づく予測を高速かつ高精度に行うことが課題である。
【解決手段】　類似事例抽出条件決定部１２は、既知事例集合Ａを用いて最適な類似事例抽出条件を自動的に決定し、予測部１３は、その条件を用いて、既知事例集合Ａから未知事例集合Ｕの未知フィールドを予測する。このとき、予測部１３は、既知事例集合Ａにおける未知フィールドの値の分布に依存する類似度を計算し、類似度に基づいて類似事例集合を抽出する。また、予測部１３は、類似度が所定の条件を満たさないことが確定した時点で、類似度計算を中止する。

予測装置の構成図



【特許請求の範囲】

【請求項１】　類似事例に基づく予測を行う予測装置であって、

１つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する１つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定する決定手段と、

前記決定手段により決定された類似事例抽出条件を用いて、前記１つ以上の類似事例データを抽出し、該１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を出力する予測手段とを備えることを特徴とする予測装置。

【請求項２】　前記決定手段は、複数の類似事例抽出条件を評価し、評価結果に基づいて、適切な類似事例抽出条件を選択することを特徴とする請求項１記載の予測装置。

【請求項３】　前記決定手段は、

類似事例抽出条件の種類を指定する種類情報と指定された類似事例抽出条件の範囲を指定する範囲情報を出力する条件出力手段と、

前記条件出力手段の出力を包含するような最大条件を求める最大条件計算手段と、

前記既知事例データの集合から既知事例入力用事例データの集合と未知事例入力用事例データの集合を生成する入力用事例生成手段と、

前記最大条件に従って、前記既知事例入力用事例データの集合から前記未知事例入力用事例データに類似する類似事例データの集合を抽出する類似事例抽出手段と、

前記未知事例入力用事例データに類似する類似事例データの集合を用いて、前記類似事例抽出条件の範囲の中から適切な類似事例抽出条件を選択し、該適切な類似事例抽出条件の情報を出力する条件決定手段とを含むことを特徴とする請求項１記載の予測装置。

【請求項４】　前記条件出力手段は、類似事例数を前記種類情報として出力し、該類似事例数の範囲を前記範囲情報として出力することを特徴とする請求項３記載の予測装置。

【請求項５】　前記条件出力手段は、類似度の閾値を前記種類情報として出力し、該閾値の範囲を前記範囲情報として出力することを特徴とする請求項３記載の予測装置。

【請求項６】　前記条件出力手段は、類似事例数および類似度を含む条件式を前記種類情報として出力し、該条件式の検査範囲を前記範囲情報として出力することを特徴とする請求項３記載の予測装置。

【請求項７】　前記入力用事例生成手段は、前記既知事例データの集合を２つに分割して、一方を前記既知事例入力用事例データの集合として出力し、他方を前記未知事例入力用事例データの集合として出力することを特徴とする請求項３記載の予測装置。

【請求項８】　前記入力用事例生成手段は、前記既知事例データの集合を前記未知事例入力用事例データの集合として出力し、該既知事例データの集合から１つの未知事例入力用事例データを削除して得られた集合を、該１つの未知事例入力用事例データに対する既知事例入力用事例データの集合として出力することを特徴とする請求項３記載の予測装置。

【請求項９】　前記入力用事例生成手段は、前記既知事例データの集合から１つ以上の事例データをサンプリングして得られた集合を、前記未知事例入力用事例データの集合として出力し、該既知事例データの集合から１つの未知事例入力用事例データを削除して得られた集合を、該１つの未知事例入力用事例データに対する既知事例入力用事例データの集合として出力することを特徴とする請求項３記載の予測装置。

【請求項１０】　前記決定手段は、

前記既知事例入力用事例データの集合に前記未知事例入力用事例データと重複する事例データが含まれているとき、前記最大条件を修正して前記類似事例抽出手段に出力する最大条件修正手段と、

前記最大条件が修正されたとき、前記類似事例抽出手段から出力された前記類似事例データの集合から事例データを削除して、修正された類似事例データの集合を前記条件決定手段に出力する類似事例削除手段とをさらに含むことを特徴とする請求項３記載の予測装置。

【請求項１１】　前記入力用事例生成手段は、前記既知事例データの集合を前記未知事例入力用事例データの集合および既知事例入力用事例データの集合として出力し、前記最大条件修正手段は、類似事例数が１つ多くなるように前記最大条件を修正し、前記類似事例削除手段は、１つの未知事例入力用事例データに対する類似事例データの集合から該１つの未知事例入力用事例データと重複する事例データを削除することを特徴とする請求項１０記載の予測装置。

【請求項１２】　前記入力用事例生成手段は、前記既知事例データの集合から１つ以上の事例データをサンプリングして得られた集合を、前記未知事例入力用事例データの集合として出力し、該既知事例データの集合を前記既知事例入力用事例データの集合として出力し、前記最大条件修正手段は、類似事例数が１つ多くなるように前記最大条件を修正し、前記類似事例削除手段は、１つの未知事例入力用事例データに対する類似事例データの集合から該１つの未知事例入力用事例データと重複する事例データを削除することを特徴とする請求項１０記載の予測装置。

【請求項１３】　前記類似事例抽出手段は、記憶に基づく推論および事例に基づく推論のうちのいずれかを用いて、前記未知事例入力用事例データに類似する類似事例データの集合を抽出することを特徴とする請求項３記載の予測装置。

【請求項１４】　前記条件決定手段は、
前記類似事例抽出条件の範囲に含まれる条件を離散化して出力する条件離散化手段と、
前記未知事例入力用事例データに類似する類似事例データの集合から、離散化された条件の各々に合致する条件ごとの類似事例データの集合を抽出する条件付き類似事例抽出手段と、
前記条件ごとの類似事例データの集合を用いて、条件ごとに前記未知事例入力用事例データの予測対象フィールドの値を予測し、条件ごとの予測値を出力する予測結果生成手段と、
前記条件ごとの予測値から条件ごとの評価値を求める条件評価手段と、
前記条件ごとの評価値に基づいて、前記離散化された条件の中から前記適切な類似事例抽出条件を選択する条件選択手段とを含むことを特徴とする請求項３記載の予測装置。
【請求項１５】　前記予測結果生成手段は、記憶に基づく推論および事例に基づく推論のうちのいずれかを用いて、前記予測結果を生成することを特徴とする請求項１４記載の予測装置。
【請求項１６】　前記条件評価手段は、前記予測値がカテゴリ値であるとき、該予測値と前記予測対象フィールドの真の値を比較し、該予測値と真の値が一致するかどうかに従って前記評価値を生成することを特徴とする請求項１４記載の予測装置。
【請求項１７】　前記条件評価手段は、前記予測値が連続値であるとき、該予測値と前記予測対象フィールドの真の値を比較し、該予測値と真の値の差を用いて前記評価値を生成することを特徴とする請求項１４記載の予測装置。
【請求項１８】　前記条件評価手段は、前記予測値に付随する確信度を加味して前記評価値を生成することを特徴とする請求項１４記載の予測装置。
【請求項１９】　前記条件評価手段は、前記予測値と前記予測対象フィールドの真の値から重みを求め、該重みを加味して前記評価値を生成することを特徴とする請求項１４記載の予測装置。
【請求項２０】　前記条件評価手段は、１つの離散化された条件と該１つの離散化された条件に合致する類似事例データの集合の少なくとも一方を用いて、類似事例抽出の実行時間を推定し、推定された実行時間を加味して前記評価値を生成することを特徴とする請求項１４記載の予測装置。
【請求項２１】　前記条件選択手段は、与えられた評価値のうち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項１４記載の予測装置。
【請求項２２】　前記条件選択手段は、与えられた評価値について移動平均を計算し、得られた平均評価値のうち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項１４記載の予測装置。
【請求項２３】　前記条件選択手段は、与えられた評価値を条件の関数により近似し、得られた近似評価値のうち最良の値に対応する条件を前記適切な類似事例抽出条件として選択することを特徴とする請求項１４記載の予測装置。
【請求項２４】　類似事例に基づく予測を行う予測装置であって、
１つ以上のフィールドからなる既知事例データの集合から、類似度に基づいて未知事例データに類似する１つ以上の類似事例データを抽出する類似事例抽出手段と、
前記１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を出力する予測結果生成手段と、
前記既知事例データの集合における前記未知フィールドの値の分布と前記未知事例データの未知フィールドの値に依存する重みをフィールドごとに計算し、フィールドごとの重みを用いて前記類似度を計算する類似度計算手段とを備えることを特徴とする予測装置。
【請求項２５】　前記類似度計算手段は、フィールドごとに前記既知事例データの集合を該フィールドの値により部分集合に分割し、前記未知事例データが属する部分集合の未知フィールドの値の分布が全体の未知フィールドの値の分布に近いほど小さくなり、該未知事例データが属する部分集合の未知フィールドの値の分布が該全体の未知フィールドの値の分布から遠いほど大きくなるような重みを計算することを特徴とする請求項２４記載の予測装置。
【請求項２６】　類似事例に基づく予測を行う予測装置であって、
類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算する類似度条件計算手段と、
前記既知事例データと未知事例データの類似度を計算し、該類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして出力し、該類似度が該類似度条件を満たさないことが確定したとき、計算を中止する条件付き類似度計算手段と、
前記条件付き類似度計算手段から出力された類似事例データを用いて、新たな類似事例データの集合を生成する生成手段とを備えることを特徴とする予測装置。
【請求項２７】　前記生成手段は、前記条件付き類似度計算手段から出力された類似事例データを前記暫定的な類似事例データの集合に加え、前記類似事例抽出条件に合致するように余分な事例データを取り除いて、新しい類似事例データの集合を生成する類似事例集合更新手段と、該新しい類似事例データの集合を暫定的な類似事例

データの集合として記憶する類似事例集合記憶手段とを含み、与えられたすべての既知事例データの処理が終了したとき、該類似事例集合記憶手段に記憶された類似事例データの集合を出力することを特徴とする請求項２６記載の予測装置。

【請求項２８】　類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、

１つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する１つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定するステップと、

決定された類似事例抽出条件を用いて、前記１つ以上の類似事例データを抽出するステップと、

前記１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項２９】　類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、

１つ以上のフィールドからなる既知事例データの集合において、未知事例データの未知フィールドに対応するフィールドの値の分布を計算するステップと、

前記分布に依存する重みをフィールドごとに計算するステップと、

フィールドごとの重みを用いて類似度を計算するステップと、

前記既知事例データの集合から、前記類似度に基づいて前記未知事例データに類似する１つ以上の類似事例データを抽出するステップと、

前記１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測し、予測値を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３０】　類似事例に基づく予測を行うコンピュータのためのプログラムを記録した記録媒体であって、

類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算するステップと、

前記既知事例データと未知事例データの類似度を計算するステップと、

前記類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして決定するステップと、

前記類似度が前記類似度条件を満たさないことが確定したとき、計算を中止するステップと、

決定された類似事例データを用いて、新たな類似事例データの集合を生成するステップとを含む処理を前記コンピュータに実行させるためのプログラムを記録したコンピュータ読み取り可能な記録媒体。

【請求項３１】　コンピュータを用いた類似事例に基づく予測方法であって、

１つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する１つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定し、

決定された類似事例抽出条件を用いて、前記１つ以上の類似事例データを抽出し、

前記１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【請求項３２】　コンピュータを用いた類似事例に基づく予測方法であって、

１つ以上のフィールドからなる既知事例データの集合において、未知事例データの未知フィールドに対応するフィールドの値の分布を計算し、

前記分布に依存する重みをフィールドごとに計算し、

フィールドごとの重みを用いて類似度を計算し、

前記既知事例データの集合から、前記類似度に基づいて前記未知事例データに類似する１つ以上の類似事例データを抽出し、

前記１つ以上の類似事例データを用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【請求項３３】　コンピュータを用いた類似事例に基づく予測方法であって、

類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを該暫定的な類似事例データの集合に加えるための類似度条件を計算し、

前記既知事例データと未知事例データの類似度を計算し、

前記類似度が前記類似度条件を満たすとき、該既知事例データを類似事例データとして決定し、

前記類似度が前記類似度条件を満たさないことが確定したとき、計算を中止し、

決定された類似事例データを用いて、新たな類似事例データの集合を生成し、

与えられたすべての既知事例データの処理が終了したとき、生成された類似事例データの集合を用いて、前記未知事例データの未知フィールドの値を予測することを特徴とする予測方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、人工知能や統計処理等の分野において、与えられた未知事例に類似する類似事例を既知事例の中から抽出し、抽出された類似事例を用いて未知事例に関する予測を行う予測装置およびその方法に関する。

【０００２】

【従来の技術】与えられた事例がどのクラス（カテゴリ）に属するかを決定するデータ分類方法は、人工知能や統計処理等の分野においてよく研究されている。事例（caseまたはinstance）とは、例えば、レコードのようなデータ構造に対応し、１つ以上のフィールドから構成される。そして、各フィールドには、属性データが記録される。

【０００３】例えば、セールスキャンペーンにおいてダイレクトメールの送り先を決定する場合、送り先となる個人の情報を事例として扱い、多数の個人をフィールドの値により分類することが考えられる。このとき、各事例には、性別、年齢、住所、職業等の属性データのフィールドが設けられる。

【０００４】また、既知事例の集合から未知事例の集合のクラスフィールドを予測する方法の１つとして、記憶に基づく推論（memory-based reasoning）、事例に基づく推論（case-based reasoning）等の類似事例に基づく予測が挙げられる。ここで、クラスフィールドは、予測対象となるフィールドを指し、既知事例は、クラスフィールドの値が既知の事例を表し、未知事例は、クラスフィールドの値が未知の事例を表す。

【０００５】上述のダイレクトメールの例では、例えば、個人がダイレクトメールに対して応答を返したかどうかを表すフラグをクラスフィールドとして設定する。そして、既にダイレクトメールを送付した個人の集合を既知事例集合とし、これからダイレクトメールを送付しようとする個人の集合を未知事例集合として、未知事例集合のクラスフィールドの値（クラス値）を予測する。これにより、応答を返す可能性の高い個人を送り先として選択することが可能になる。

【０００６】未知事例のクラスフィールドを予測するためには、まず、事例間の類似度を定義し、その未知事例に類似しているいくつかの事例を既知事例集合から取り出す。そして、それらの類似事例を基にして、未知事例のクラスフィールドを予測する。この予測方法は、ニューラルネットワークや決定木等を用いた学習やモデル化に基づく他の予測方法とは異なり、学習やモデル化のための時間が不要で、既知事例の増加に柔軟に対応できるという利点を持っている。

【０００７】

【発明が解決しようとする課題】しかしながら、上述した従来の予測方法には、次のような問題がある。

【０００８】ある未知事例に関する予測を行うためには、その未知事例と各既知事例との間の類似度を計算して、類似事例を抽出しなければならない。実は、類似事例に基づく予測に要する処理時間のほとんどが、この類似度計算の時間である。従来の予測方法では、類似度計算の時間を短縮するために、既知事例に対して一定の前処理を行う必要があった（毛利隆夫(Takao Mori)，「Nearest Neighbor法と記憶に基づく推論」，人口知能学会誌，Vol.12 No.2，p188-p195，March 1997.）。

【０００９】この前処理では、既知事例集合をグループ分けして、未知事例との類似度が低いと思われるグループを類似度計算の対象から外す処理や、既知事例集合から不要な事例を削除する処理等が行われる。しかし、このような前処理を行うと、既知事例の増加に柔軟に対応することが難しくなる。

【００１０】また、従来は、抽出される類似事例に関する条件として、類似度の高い上位２０個の事例というようなデフォルトの条件をそのまま用いるか、あるいはユーザの手によって直接入力された条件を用いていた。しかし、どの条件が最適であるかは事例集合の性質および問題設定の目的によって大きく異なることが多く、適当な条件を指定しなければユーザが満足する結果が得られないことも多い。そこで、目的に合った条件を選ぶために、ユーザが条件を変更しながら類似事例の抽出を複数回実行し、その結果を評価して、最適な条件を選ぶことが多かった。

【００１１】また、従来の予測方法では、類似事例を選択する際、クラス値に対する影響度に応じてフィールド毎に重みを設定することで、精度の向上が図られている。自動的な重み設定の方法としては、統計情報を用いたCross Category Feature（ＣＣＦ）法が存在する。しかし、ＣＣＦ法によれば、既知事例集合のクラス値の分布に無関係に重みが設定され、クラス値の分布の変化による重みへの影響が十分ではない。このため、既知事例集合のクラス値の分布が大きく偏っている場合等には、十分な精度が出ないことが多かった。

【００１２】このように、従来の予測方法では、既知事例集合に対する前処理なしに予測の高速化を行うことができず、既知事例の増加に対する柔軟性が損なわれるという問題がある。また、十分な予測精度を得るために類似事例の抽出を複数回行うので、ユーザの負担が大きく、実行回数に応じて計算時間が増大するという問題もある。さらに、既知事例集合のクラス値の分布が大きく偏っている場合等、問題の性質によっては、ＣＣＦ法により十分な精度が得られないという問題もある。

【００１３】本発明の課題は、類似事例に基づいて未知事例に関する予測を高速かつ高精度に行う予測装置およびその方法を提供することである。

【００１４】

【課題を解決するための手段】図１は、本発明の予測装置の原理図である。

【００１５】本発明の第１の原理による予測装置は、決定手段１と予測手段２を備え、類似事例に基づく予測を行う。

【００１６】決定手段１は、１つ以上のフィールドからなる既知事例データの集合から未知事例データに類似する１つ以上の類似事例データを抽出するための類似事例抽出条件を自動的に決定する。

【００１７】予測手段２は、決定手段１により決定された類似事例抽出条件を用いて、上記１つ以上の類似事例データを抽出し、それらの類似事例データを用いて、未知事例データの未知フィールドの値を予測し、予測値を出力する。

【００１８】決定手段１は、既知事例データの集合から疑似的な未知事例データ（未知事例入力用事例データ）の集合を生成し、各未知事例入力用事例データのクラスフィールドが未知であるものとして予測を行う。そして、例えば、最良の結果が得られるような類似事例抽出条件を、最適な類似事例抽出条件として決定する。

【００１９】また、予測手段２は、決定された類似事例抽出条件を用いて未知事例データの未知フィールド（クラスフィールド）の予測を行い、その類似事例抽出条件に応じた精度の予測値を得る。

【００２０】このように、既知事例データの集合のみを用いてあらかじめ疑似的な予測処理を行うことで、最適な類似事例抽出条件が自動的に決定されるので、実際の未知事例データの予測は１回だけ行えばよい。したがって、従来のように未知事例データの予測を何度も繰り返さなくても、精度の高い予測値が得られ、処理が高速化される。

【００２１】また、本発明の第２の原理による予測装置は、類似事例抽出手段３、予測結果生成手段４、および類似度計算手段５を備え、類似事例に基づく予測を行う。

【００２２】類似事例抽出手段３は、１つ以上のフィールドからなる既知事例データの集合から、類似度に基づいて未知事例データに類似する１つ以上の類似事例データを抽出する。

【００２３】予測結果生成手段４は、上記１つ以上の類似事例データを用いて、未知事例データの未知フィールドの値を予測し、予測値を出力する。

【００２４】類似度計算手段５は、上記既知事例データの集合における上記未知フィールドの値の分布と上記未知事例データの未知フィールドの値に依存する重みをフィールドごとに計算し、フィールドごとの重みを用いて類似度を計算する。

【００２５】類似度計算手段５が計算するフィールドごとの重みは、既知事例データの集合のクラス値の分布の影響を受けるため、それを用いて計算される類似度もそのクラス値の分布の影響を受ける。類似事例抽出手段３は、このような類似度に基づいて類似事例データを抽出し、予測結果生成手段４は、それらの類似事例データを用いて未知事例データのクラス値を予測する。

【００２６】これにより、既知事例データの集合のクラス値の分布の偏りが大きい場合等の、従来のＣＣＦ法では十分な精度が得られないような問題においても、高精度な予測が行えるようになる。

【００２７】また、本発明の第３の原理による予測装置は、類似度条件計算手段６、条件付き類似度計算手段７、および生成手段８を備え、類似事例に基づく予測を行う。

【００２８】類似度条件計算手段６は、類似事例抽出条件と既に得られている暫定的な類似事例データの集合を用いて、既知事例データを暫定的な類似事例データの集合に加えるための類似度条件を計算する。

【００２９】条件付き類似度計算手段７は、既知事例データと未知事例データの類似度を計算し、類似度が類似度条件を満たすとき、その既知事例データを類似事例データとして出力し、類似度が類似度条件を満たさないことが確定したとき、計算を中止する。

【００３０】生成手段８は、条件付き類似度計算手段７から出力された類似事例データを用いて、新たな類似事例データの集合を生成する。

【００３１】条件付き類似度計算手段７は、類似度計算の途中で、類似度条件計算手段６が出力した類似度条件が満たされないことが確定した時点で、その既知事例データは類似事例データになり得ないと判断し、計算を中止する。そして、生成手段８は、類似度計算が中止されなかった既知事例データを用いて類似事例データの集合を生成し、予測装置は、その類似事例データの集合を用いて予測を行う。

【００３２】このように、類似事例データの抽出において多大な時間を要する類似度計算を途中で打ち切ることで、不要な計算時間が削減され、類似事例の抽出が高速化される。したがって、既知事例データの集合に対する前処理を行わなくても、高速な予測を行うことが可能になる。

【００３３】例えば、図１の決定手段１と予測手段２は、それぞれ、後述する図２の類似事例抽出条件決定部１２と予測部１３に対応し、図１の類似事例抽出手段３と類似度計算手段５は、後述する図５の類似事例抽出部４１に対応し、図１の予測結果生成手段４は、図５の予測結果生成部４２に対応する。また、例えば、図１の類似度条件計算手段６と条件付き類似度計算手段７は、それぞれ、後述する図６の類似度条件計算部５４と条件付き類似度計算部５１に対応し、図１の生成手段８は、図６の類似事例集合更新部５２と類似事例集合記憶部５３に対応する。

【００３４】

【発明の実施の形態】以下、図面を参照しながら、本発明の実施の形態を詳細に説明する。

【００３５】本発明では、事例間の類似度計算の際に、ある既知事例が類似事例になり得ないことが分かった時点で計算を中止する。これにより、類似事例の抽出が高速化され、既知事例集合に対する前処理を行わなくても、予測が高速化される。

【００３６】また、本発明では、従来、ユーザが明示的に処理を複数回実行することで得ていた最適な類似事例

抽出条件を、１回の実行で自動的に得ることができるようにする。ここでは、既知事例集合またはその部分集合をテスト用の疑似的な未知事例集合として用い、各未知事例のクラスフィールドが未知であるものとして予測を行う。そして、最良の結果が得られるような類似事例抽出条件を、最適な類似事例抽出条件として決定する。

【００３７】このとき、条件を変更しながら類似事例の抽出を複数回行うことを避けるために、すべての類似事例抽出条件を包含する最も広い条件をあらかじめ計算しておき、その条件を用いて類似事例の抽出を１回だけ行う。その後、得られた類似事例を評価することで、最適な類似事例抽出条件を決定する。これにより、予測精度を損なうことなく、処理を高速化することができる。

【００３８】また、本発明では、既知事例集合のクラス値の分布の影響を受け、クラス値の分布の変化による重みへの影響が従来のＣＣＦ法よりも大きくなるような影響度計算方法を用いる。これにより、既知事例集合のクラス値の分布の偏りが大きい場合等の、ＣＣＦ法では十分な精度が得られないような問題においても、高精度な予測が行えるようになる。

【００３９】図２は、本発明の予測装置の構成図である。図２の予測装置１１は、例えば、コンピュータを用いて構成され、類似事例抽出条件決定部１２と予測部１３を備える。類似事例抽出条件決定部１２および予測部１３の機能は、例えば、コンピュータのメモリに格納されたプログラムを実行することで実現される。

【００４０】類似事例抽出条件決定部１２は、既知事例集合Ａを用いて最適な類似事例抽出条件を決定し、それを出力する。予測部１３は、類似事例抽出条件決定部１２の出力を類似事例抽出条件として用い、既知事例集合Ａから未知事例集合Ｕのクラスフィールドを予測して、予測結果を出力する。

【００４１】図３は、図２の類似事例抽出条件決定部１２の構成図である。図３の類似事例抽出条件決定部１２は、入力用事例生成部２１、類似事例抽出部２２、類似事例削除部２３、最適条件決定部２４、条件出力部２５、最大条件計算部２６、および最大条件修正部２７を備える。

【００４２】入力用事例生成部２１は、既知事例集合Ａから既知事例入力用事例集合Ｂと未知事例入力用事例集合Ｃの２つの事例集合を生成する。条件出力部２５は、あらかじめ保持している類似事例抽出条件の種類と最適条件選択範囲を出力する。最適条件選択範囲は、良好な予測を行うために最適な類似事例抽出条件を選択する際の条件の範囲を表す。

【００４３】最大条件計算部２６は、条件出力部２５が出力するすべての条件を包含するような最も広い条件を求め、それを最大条件として出力する。最大条件修正部２７は、既知事例入力用事例集合Ｂに未知事例入力用事例集合Ｃと重複した事例が含まれている場合に、最大条件を修正して出力し、それ以外の場合は最大条件をそのまま出力する。

【００４４】類似事例抽出部２２は、入力用事例生成部２１の出力Ｂ、Ｃを入力とし、最大条件修正部２７が出力する条件に合致する類似事例集合を出力する。類似事例削除部２３は、入力用事例生成部２１の出力である既知事例入力用事例集合Ｂに重複した事例が含まれている場合に、類似事例抽出部２２の出力である類似事例集合から事例を削除して、類似事例集合を修正する。

【００４５】最適条件決定部２４は、類似事例削除部２３の出力を評価し、条件出力部２５の出力である最適条件選択範囲の中から、良好な予測を行うために最適な類似事例抽出条件を決定し、それを出力する。

【００４６】図４は、図３の最適条件決定部２４の構成図である。図４の最適条件決定部２４は、条件離散化部３１、条件付き類似事例抽出部３２、予測結果生成部３３、条件評価部３４、および最適条件選択部３５を備える。

【００４７】条件離散化部３１は、条件出力部２５の出力である最適条件選択範囲を離散化して出力する。条件付き類似事例抽出部３２は、類似事例削除部２３の出力から、条件離散化部３１の出力の条件ごとに合致する類似事例集合を抽出する。

【００４８】予測結果生成部３３は、条件付き類似事例抽出部３２の出力を用いて、条件離散化部３１の出力の条件ごとに予測を行う。条件評価部３４は、予測結果生成部３３の出力である予測結果を評価し、条件離散化部３１の出力の条件ごとに評価値を求める。最適条件選択部３５は、条件評価部３４の出力である評価値に基づいて、良好な予測を行うために最適な類似事例抽出条件を選択する。

【００４９】このような構成によれば、類似事例抽出部２２による類似事例の抽出は１回行われるだけであり、最適な類似事例抽出条件は最適条件決定部２４により自動的に決定される。類似事例に基づく予測では、計算時間のほとんどが類似事例抽出の際の類似度計算に費やされるため、類似事例抽出を１回で済ますことにより、ユーザが明示的に複数回実行する場合と比べて、計算時間を大幅に短縮できる。

【００５０】ところで、図３の条件出力部２５は、類似事例抽出条件の種類として、例えば、１つの未知事例に対する類似事例の数がｋであるという条件を出力し、最適条件選択範囲としてｋの値の検査範囲を出力する。ｋの値の範囲としては、例えば、$1 \leq k \leq k1$なる整数が指定される。

【００５１】この場合、最大条件計算部２６は、類似事例数がｋ１であるという条件を最大条件として出力する。ｋ１の値は、ユーザが指定することもでき、ｋ１＝（既知事例集合Ａの事例数の平方根）のように、システムが自動的に設定することもできる。

【００５２】また、条件出力部２５は、類似事例抽出条件の種類として、類似度が閾値ｓ以上という条件を出力し、最適条件選択範囲として閾値ｓの値の検査範囲を出力することもできる。ｓの値の範囲としては、例えば、ｓ１≦ｓ≦ｓ２が指定される。この場合、最大条件計算部２６は、類似度がｓ１以上という条件を最大条件として出力する。

【００５３】ｓ１、ｓ２の値は、それぞれ、ユーザが指定することもでき、システムが自動的に設定することもできる。後者の場合、例えば、$\alpha=1$、$\beta=100$のよ

$$a=0 \quad or \quad 1/k1 \leqq a \leqq 1$$

$$b=0 \quad or \quad s1 \leqq b \leqq s2$$

ただし、ａ＞０もしくはｂ＞０であるものとする。ｋ１、ｓ１、およびｓ２の値は、上述したような方法で設

$$s \geqq s1 \text{ もしくは } k=k1$$

次に、図５は、図２の予測部１３の構成図である。図５の予測部１３は、類似事例抽出部４１と予測結果生成部４２を備える。類似事例抽出部４１は、既知事例集合Ａと未知事例集合Ｕを入力とし、類似事例抽出条件決定部１２が出力する条件に合致する類似事例集合を出力する。予測結果生成部４２は、類似事例抽出部４１の出力を用いて予測を行い、予測結果を出力する。

【００５５】図６は、図５の類似事例抽出部４１の構成図である。図６の類似事例抽出部４１は、条件付き類似度計算部５１、類似事例集合更新部５２、類似事例集合記憶部５３、および類似度条件計算部５４を備え、未知事例集合Ｕの各事例ごとに、既知事例集合Ａから類似事例集合を抽出する。

【００５６】条件付き類似度計算部５１は、既知事例集合Ａから既知事例を１つずつ取り出して、既知事例と未知事例の類似度を計算し、与えられた類似度条件を満たす既知事例とその類似度を出力する。ただし、与えられた類似度条件を満たさないことが分かった時点で、計算を中止する。

【００５７】類似事例集合更新部５２は、既に得られている暫定的な類似事例集合に条件付き類似度計算部５１の出力を加え、類似事例抽出条件に合致するように、余分な事例を取り除き、新しい類似事例集合を出力する。類似事例集合記憶部５３は、類似事例集合更新部５２の出力を現在の暫定的な類似事例集合として記憶し、それを類似事例集合更新部５２および類似度条件計算部５４に出力する。

【００５８】類似度条件計算部５４は、類似事例集合記憶部５３の内容と類似事例抽出条件から、ある事例が新しく類似事例集合に加わるための必要十分条件である類似度条件を計算し、それを条件付き類似度計算部５１に出力する。

【００５９】このように、条件付き類似度計算部５１が類似度条件に基づいて不要な類似度計算を途中で中止することにより、既知事例集合からの類似事例抽出が高速うにあらかじめ設定されたパラメータを用いて、ｓ１＝α／（既知事例集合Ａのフィールド数）、ｓ２＝β／（既知事例集合Ａのフィールド数）のように設定することができる。

【００５４】また、条件出力部２５は、類似事例抽出条件の種類として、類似事例数ｋと類似度ｓを含む条件を出力することもできる。例えば、ａｋ＋ｂ／ｓ≦１という類似事例抽出条件を出力する場合、次のようなパラメータａ、ｂの値の検査範囲を最適条件選択範囲として出力する。

（１）

（２）

定することができる。この場合、最大条件計算部２６は、次のような条件を最大条件として出力する。

（３）

化され、類似事例に基づく予測を高速化することができる。この類似事例抽出部４１の構成を図３の類似事例抽出部２２に実装して、類似事例抽出条件決定部１２の処理をより高速化することもできる。

【００６０】次に、図７から図４９までを参照しながら、上述した予測装置１１の動作についてより詳細に説明する。

【００６１】図７は、図３の入力用事例生成部２１の第１の例を示している。図７の入力用事例生成部２１は分割部６１を備え、既知事例集合Ａを２つに分割して、一方を既知事例入力用事例集合Ｂとして出力し、もう一方を未知事例入力用事例集合Ｃとして出力する。分割部６１における分割方法としては、例えば、ランダムサンプリング等が考えられる。

【００６２】このとき、最適な類似事例抽出条件が既知事例集合Ａと既知事例入力用事例集合Ｂの間で大きく異ならないようにするために、未知事例入力用事例集合Ｃの事例数が既知事例集合Ａの事例数に比べて十分小さいことが望ましい。この構成では、既知事例入力用事例集合Ｂには、未知事例と重複する事例は含まれないため、最大条件修正部２７および類似事例削除部２３は何も行わず、入力をそのまま出力する。

【００６３】図８は、図３の入力用事例生成部２１の第２の例を示している。図８の入力用事例生成部２１は事例削除部６２を備え、既知事例集合Ａをそのまま未知事例入力用事例集合Ｃとして出力する。事例削除部６２は、未知事例入力用事例集合Ｃの各未知事例ごとに、既知事例集合Ａからその未知事例を削除する。そして、未知事例ごとに異なる事例集合を生成し、それを既知事例入力用事例集合Ｂとする。

【００６４】この場合、既知事例入力用事例集合Ｂは、未知事例を含まず、かつ、既知事例集合Ａに最も近い事例集合であるといえる。この構成においても、既知事例入力用事例集合Ｂには、未知事例と重複する事例は含まれないため、最大条件修正部２７および類似事例削除部

２３は何も行わず、入力をそのまま出力する。

【００６５】図９は、図３の入力用事例生成部２１の第３の例を示している。図９の入力用事例生成部２１は、既知事例集合Ａをそのまま未知事例入力用事例集合Ｃとして出力し、同一の既知事例集合Ａをそのまま既知事例入力用事例集合Ｂとしても出力する。

【００６６】この構成では、既知事例入力用事例集合Ｂに、未知事例と重複する事例が含まれるため、最大条件修正部２７は、類似事例を１つ余計に抽出できるように最大条件を修正し、類似事例削除部２３は、各未知事例の類似事例集合からその未知事例と重複する事例を削除する。

【００６７】図１０は、図３の入力用事例生成部２１の第４の例を示している。図１０の入力用事例生成部２１は、図８の入力用事例生成部２１にサンプリング部６３を付加した構成を持つ。サンプリング部６３は、既知事例集合Ａの事例数が一定数（例えば、１０００）よりも多い場合に、ランダムサンプリング等により事例をサンプリングする。そして、既知事例集合Ａの一部を未知事例入力用事例集合Ｃとして出力する。

【００６８】事例削除部６２は、サンプリング部６３が出力した未知事例入力用事例集合Ｃの各未知事例ごとに、既知事例集合Ａからその未知事例を削除する。そして、未知事例ごとに異なる事例集合を生成し、それを既知事例入力用事例集合Ｂとする。

【００６９】この構成によれば、既知事例集合Ａが比較的大きな場合に、未知事例入力用事例集合Ｃの大きさを限定することができ、後続する類似事例抽出部２２の処理を高速化することができる。

【００７０】図１１は、図３の入力用事例生成部２１の第５の例を示している。図１１の入力用事例生成部２１は、図９の入力用事例生成部２１にサンプリング部６３を付加した構成を持つ。サンプリング部６３の動作は、図１０の場合と同様である。

【００７１】この構成では、既知事例入力用事例集合Ｂに、未知事例と重複する事例が含まれるため、最大条件修正部２７は、図９の場合と同様に最大条件を修正し、類似事例削除部２３は、各未知事例の類似事例集合からその未知事例と重複する事例を削除する。

【００７２】ここで、具体的な事例集合を用いて入力用事例生成部２１の処理を説明する。例えば、ダイレクトメールの送り先からの応答の有無を予測するために、図１２のような既知事例集合Ａが与えられたものとする。

【００７３】図１２において、１つの行が１つの事例のレコードに対応し、“氏名”は、事例を識別するための個人名を表す。この既知事例集合には、“Ａ”、“Ｂ”、“Ｃ”、“Ｄ”、“Ｅ”、“Ｆ”、および“Ｇ”の７人の事例が含まれている。各事例は、“年齢”、“性別”、“職業”、“結婚”、および“応答”のフィールドを含んでおり、このうち、“応答”がクラスフィールドに対応する。

【００７４】例えば、“Ａ”さんの事例には、年齢が“３０歳”で、性別が“男”で、職業が“公務員”で、結婚については“既婚”で、ダイレクトメールへの応答が“あり”であることが記録されている。また、“Ｃ”さんの事例には、年齢が“４０歳”で、性別が“女”で、職業が“無職”で、結婚については“既婚”で、ダイレクトメールへの応答が“なし”であることが記録されている。

【００７５】ここで、図１１に示した入力用事例生成部２１を用いてサンプリングを行い、図１３のような未知事例入力用事例集合Ｃが得られたとする。図１３の未知事例入力用事例集合には、“Ａ”、“Ｃ”、“Ｅ”、および“Ｇ”の４人の事例が含まれている。この場合、既知事例入力用事例集合Ｂとしては、図１２の事例集合がそのまま用いられる。

【００７６】図３の類似事例抽出部２２は、記憶に基づく推論もしくは事例に基づく推論により、入力用事例生成部２１が出力した集合から類似事例を抽出する。記憶に基づく推論および事例に基づく推論において、類似事例抽出部２２は基本的に同じ処理を行う。以下に類似事例抽出部２２による類似度計算の一例を示す。

【００７７】まず、事例の各フィールドごとに、クラス値の決定に対する影響度を計算する。ここでは、統計情報を基にした影響度計算方法として広く知られているＣＣＦ法を用いることにする。この方法では、クラスフィールドも含めて、各フィールドごとに、フィールド値がいくつかの領域に分けられる。

【００７８】例えば、図１２の“性別”、“職業”、“結婚”、および“応答”のように、カテゴリを表すカテゴリ値フィールドであれば、そのフィールドが表現し得るカテゴリごとにフィールド値を分類し、図１２の“年齢”のように数値を表す数値フィールドであれば、数値の区間ごとにフィールド値を分類する。ただし、“年齢”のフィールドは、離散的な数値を表すため、カテゴリ値フィールドとして扱うことも可能である。

【００７９】今、既知事例集合の事例番号をｉとし、フィールド番号をｊとすると、フィールドｊの値が領域ｖに含まれているときに、クラスフィールドの値が領域ｃに含まれている条件付き確率を、ｐ（ｊ，ｖ，ｃ）のように表すことができる。このとき、フィールドｊの重みｗ（ｊ，ｖ）は、次式で与えられる。

【００８０】

【数１】

$$w(j,v)=\sum_{c} p(j,v,c)^2 \qquad (4)$$

【００８１】次に、２つのフィールド値間の距離を定義する。ここでは、一例として、最も単純なフィールド間距離の１つである次式のような距離ｄ（ｊ）を用いる。

【００８２】

【数2】

$$d(j)=\begin{cases}\dfrac{\text{field値間の差}}{\text{標準偏差}} & (\text{数値 field})\\ \begin{cases}0 & (\text{field値が一致})\\ 1 & (\text{field値が不一致})\end{cases} & (\text{カテゴリ値 field})\end{cases} \qquad (5)$$

【００８３】（５）式によれば、数値フィールドの場合は、フィールド値の分布から求められた標準偏差を分母とし、２つのフィールド値の差を分子とする値が、距離ｄ（ｊ）として用いられる。また、カテゴリ値フィールドの場合は、２つのフィールド値が一致したときｄ（ｊ）＝０と定義され、それが一致しなかったときｄ（ｊ）＝１と定義される。

【００８４】次に、２つの事例間の類似度を定義する。ここでは、一例として、最も単純な類似度の１つである次式のような類似度Ｓを用いる。

【００８５】

【数3】

$$S = 1\Big/\sqrt{\sum_j w(j,v(j))\times d(j)^2} \qquad (6)$$

【００８６】ただし、ｖ（ｊ）は、未知事例のフィールドｊの値が属している領域を表す。こうして、（４）式、（５）式、および（６）式により事例間類似度が定義されたので、各未知事例についてすべての既知事例との類似度を計算し、最大条件修正部２７の出力条件に合致するような既知事例を選択することで、未知事例ごとに類似事例を抽出することができる。また、このような類似度計算は、図６の条件付き類似度計算部５１にも適用することができる。

【００８７】例えば、条件出力部２５が類似事例の数ｋを類似事例抽出条件として出力し、ｋ≦５を最適条件選択範囲として出力したとすると、最大条件計算部２６は、ｋ＝５を最大条件として出力する。このとき、最大条件修正部２７は、類似事例を１つ余計に抽出できるように、最大条件をｋ＝６に修正し、類似事例抽出部２２は、修正された最大条件に従って類似事例を抽出する。

【００８８】このとき、図１２および図１３に示した事例集合からは、例えば、図１４、１５、１６、および１７に示すような類似事例集合が生成される。図１４、１５、１６、および１７の類似事例集合は、それぞれ、テスト用の未知事例“Ａ”、“Ｃ”、“Ｅ”、および“Ｇ”に類似する事例の集合を表す。

【００８９】ここでは、いずれの類似事例集合も、修正された最大条件に対応する６つの類似事例から構成され、それらの類似事例は類似度の大きい順に並べられている。各類似事例集合は、比較対象の未知事例と同じ類似事例を含んでおり、その類似事例の類似度は“＊＊＊”で表されている。

【００９０】類似事例削除部２３は、これらの類似事例集合から対応する未知事例と重複する事例を削除して、類似事例集合を修正する。図１４、１５、１６、および１７の類似事例集合から重複事例がそれぞれ削除された結果、図１８、１９、２０、および２１に示すような類似事例集合が生成される。例えば、図１４の類似事例集合からは事例“Ａ”が削除されて、図１８の類似事例集合が生成されている。修正された類似事例集合は、それぞれ、修正される前の最大条件に対応する５つの類似事例から構成されている。

【００９１】こうして修正された類似事例集合は、図４の最適条件決定部２４に入力され、最適な類似事例抽出条件を決定するために利用される。図４の条件離散化部３１は、連続な条件をあらかじめ決められた方法で離散化する。最も簡単な離散化方法としては、あらかじめ離散化数を決めておき、等間隔で条件を離散化する方法が挙げられる。

【００９２】例えば、類似度がｓ以上という類似事例抽出条件と、ｓ１≦ｓ≦ｓ２という最適条件選択範囲が与えられた場合、次式のような類似度ｓ（ｉ）を用いて、条件を離散化することができる。

$$s(i)=s1+i*(s2-s1)/100 \qquad (7)$$

ただし、ｉは、０≦ｉ≦１００なる整数である。（７）式により、ｓは、ｓ（０）からｓ（１００）までの１０１点に離散化され、これに対応して１０１個の類似事例抽出条件が生成される。

【００９３】また、類似事例数もしくは類似事例数と類似度を含む条件が類似事例抽出条件として与えられた場合も、同様の方法により条件を離散化することができる。例えば、類似事例数がｋという類似事例抽出条件と、ｋ≦５という最適条件選択範囲が与えられた場合、類似事例抽出条件は、“類似事例数＝１”、“類似事例数＝２”、“類似事例数＝３”、“類似事例数＝４”、および“類似事例数＝５”の５つの条件に離散化される。

【００９４】次に、条件付き類似事例抽出部３２は、入力された類似事例集合から離散化された各条件に合致する類似事例を抽出し、条件ごとに異なる類似事例集合を出力する。

【００９５】例えば、図１８、１９、２０、および２１に示した類似事例集合が入力された場合、“類似事例数＝１”という条件に従うと、各類似事例集合から類似度

が最大の類似事例が１つずつ抽出され、図２２、２３、２４、および２５に示すような類似事例集合が生成される。

【００９６】また、“類似事例数＝２”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が２つずつ抽出され、図２６、２７、２８、および２９に示すような類似事例集合が生成される。

【００９７】また、“類似事例数＝３”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が３つずつ抽出され、図３０、３１、３２、および３３に示すような類似事例集合が生成される。

【００９８】また、“類似事例数＝４”という条件に従うと、各類似事例集合から、類似度の大きなものから順に類似事例が４つずつ抽出され、図３４、３５、３６、および３７に示すような類似事例集合が生成される。

【００９９】また、“類似事例数＝５”という条件は類似事例抽出部２２が用いた条件と一致するため、この条件に従うと、図１８、１９、２０、および２１の類似事例集合がそのまま出力される。

【０１００】次に、予測結果生成部３３は、条件付き類似事例抽出部３２が出力した類似事例集合を用いて、条件ごとにクラスフィールドの予測を行う。予測結果生成部３３が記憶に基づく推論により予測を行う場合、類似事例集合および各類似事例に付随する類似度を用いて、以下のような処理を行う。

【０１０１】まず、クラス値がカテゴリ値の場合、クラス値ごとに、そのクラス値ｃを有する類似事例の事例間類似度を足し合わせることで、クラス値ごとの類似度Ｔ（ｃ）を計算する。そして、類似度Ｔ（ｃ）が最大となるクラス値を予測クラス値（予測値）ｃ（ｐｒｅｄｉｃｔ）として、予測値のもっともらしさを表す確信度Ｐを、次式により定義する。

【０１０２】

【数４】

$$P = \frac{T(c(predict))}{\sum_{c} T(c)} \qquad (8)$$

【０１０３】（８）式の分子のＴ（ｃ（ｐｒｅｄｉｃｔ））は、予測値ｃ（ｐｒｅｄｉｃｔ）に対応する類似度を表し、（８）式の分母は、すべてのクラス値についてのＴ（ｃ）の総和を表す。したがって、確信度Ｐは、最大値が１であるような正の数値である。このような予測値の計算方法のほかに、類似事例集合の事例中に最も多く現れるクラス値を予測値としてもよい。

【０１０４】また、クラス値が連続値の場合は、予測値および確信度を、例えば、次式により定義することができる。

【０１０５】

【数５】

$$c(predict) = \sum_{i}^{n} S(i)c(i) \Big/ \sum_{i}^{n} S(i) \qquad (9)$$

$$P = \frac{1}{\frac{\sum_{i}^{n} S(i)(c(i) - c(predict))^2}{\sigma(c)^2 \sum_{i}^{n} S(i)} + 1} \qquad (10)$$

【０１０６】ここで、ｎは類似事例集合に含まれる事例数を表し、Ｓ（ｉ）は事例ｉ（ｉ＝１，２，．．．，ｎ）の事例間類似度を表し、ｃ（ｉ）は事例ｉのクラス値を表し、σ（ｃ）はクラス値ｃ（ｉ）の分布における標準偏差を表す。こうして算出された予測値および確信度は、予測結果として出力される。

【０１０７】例えば、図２２から図３７に示した条件ごとの類似事例集合の場合、クラスフィールド“応答”はカテゴリフィールドであるから、クラス値ごとの類似度Ｔ（ｃ）が計算され、最大のＴ（ｃ）に対応するクラス値が予測値として求められる。その結果、図３８に示すような予測値が得られる。

【０１０８】図３８のマトリクスにおいて、各行は、未知事例入力用事例集合Ｃの１つの未知事例に対応し、各列は、１つの類似事例抽出条件（類似事例数）に対応する。したがって、マトリクスの１つの要素（セル）は、列に対応する類似事例抽出条件に従って抽出された類似事例集合を用いて、行に対応する未知事例のクラスフィールド“応答”の値を予測したときの予測値を表す。

【０１０９】例えば、“類似事例数＝１”の場合は、図２２、２３、２４、および２５に示したように、各類似事例集合は１つの事例しか含まないため、その事例の“応答”の値がそのまま予測値として採用される。

【０１１０】また、“類似事例数＝２”の場合は、図２６、２７、２８、および２９に示したように、各類似事例集合は２つの事例を含む。例えば、未知事例“Ａ”に対応する図２６の類似事例集合では、クラス値“なし”を有する事例とクラス値“あり”を有する事例が１つずつ含まれており、前者の方が類似度が大きいため、“なし”が予測値として採用される。

【０１１１】また、“類似事例数＝５”の場合は、図１８、１９、２０、および２１に示したように、各類似事例集合は５つの事例を含む。例えば、未知事例“Ａ”に対応する図１８の類似事例集合では、クラス値“なし”を有する３つの事例とクラス値“あり”を有する２つの事例が含まれており、前者の３つの事例の類似度の合計は９．５であり、後者の２つの事例の類似度の合計は７である。したがって、“なし”の類似度の方が“あり”の類似度より大きいため、“なし”が予測値として採用される。

【０１１２】次に、予測結果生成部３３が事例に基づく推論により予測を行う場合、予測結果生成部３３は図３９に示すような構成を持つ。事例に基づく推論の予測結果生成部と記憶に基づく推論の予測結果生成部との相違

点は、前者が予測修正部７２を備えていることである。

【０１１３】図３９において、一次予測生成部７１は、上述した記憶に基づく推論の予測結果生成部に対応し、予測修正部７２は、一次予測生成部７１の出力である一次予測結果を先見的知識等により修正して、予測結果生成部３３の出力を生成する。先見的知識は、過去の予測結果に基づく経験則を表す。

【０１１４】上述した記憶に基づく推論または事例に基づく推論を用いた予測結果生成部３３は、図５に示した予測部１３内の予測結果生成部４２としても用いることができる。

【０１１５】次に、条件評価部３４は、未知事例入力用事例集合Ｃを用いて条件ごとの予測結果を評価し、各条件に対する評価値を生成する。テスト用の未知事例のクラス値は以下では、評価値の値が大きいほど良好な評価結果を表すものとして、条件評価部３４のいくつかの実施形態について説明する。

【０１１６】まず、クラス値がカテゴリ値である場合、条件評価部３４は、予測結果生成部３３から出力された予測値と未知事例のクラス値（真のクラス値）を比較し、予測値と一致した事例の数を評価値とする。

【０１１７】この評価方法により、図３８に示した予測結果を評価すると、図４０に示すような評価値が得られる。ここでは、図１３の未知事例入力用事例集合の各事例のクラスフィールド“応答”の値と図３８の対応する予測値が比較される。類似事例数＝１，２，４，５の場合は、未知事例“Ｃ”および“Ｅ”についてのみ両者が一致しているので、評価値は“２”となり、類似事例数＝３の場合は、未知事例“Ａ”、“Ｃ”、および“Ｅ”について両者が一致しているので、評価値は“３”となる。

【０１１８】また、予測値と一致した事例の数から一致しなかった事例の数を差し引いた結果を評価値としてもよい。例えば、未知事例入力用事例集合Ｃの未知事例数がｍであり、一致した事例数がｍ１である場合、一致しなかった事例数はｍ－ｍ１となり、評価値はｍ１－（ｍ－ｍ１）＝２×ｍ１－ｍとなる。

【０１１９】また、クラス値が連続値である場合、条件評価部３４は、予測値と未知事例のクラス値との差の絶対値の平均をとり、得られた平均値に－１を掛けて評価値とする。この場合、評価値は負の値となる。

【０１２０】これらの評価方法では、予測結果生成部３３の出力のうち予測値のみを用いて評価値が生成されているが、さらに確信度を加味して評価値を生成することもできる。まず、クラス値がカテゴリ値の場合、条件評価部３４は、評価値の初期値を０とおき、各未知事例に対して、予測値と未知事例のクラス値が一致した場合にはその予測値の確信度を評価値に加え、両者が一致しなかった場合にはその予測値の確信度を評価値から減ずる。これにより、最終的に確信度を加味した評価値が得られる。

【０１２１】また、予測値と未知事例のクラス値を用いてクラスフィールドの重みを求め、その重みを加味して評価値を生成することもできる。例えば、クラスフィールドが“＋”と“－”の２値のいずれかをとり、ユーザは、真のクラス値が“＋”であるときに“－”という予測値を得ることはできるだけ避けたいと考えているとする。

【０１２２】このとき、ユーザは、避けたい予測パターンとして（予測値，真のクラス値）＝（－，＋）を指定し、条件評価部３４は、予測値と未知事例の真のクラス値の組合せに応じて、以下のような重みｗを設定する。

（＋，＋）の場合　ｗ＝１．０
（＋，－）の場合　ｗ＝１．０
（－，＋）の場合　ｗ＝２．０
（－，－）の場合　ｗ＝１．０

そして、評価値の初期値を０とおき、各未知事例に対して、予測値と未知事例のクラス値が一致した場合には両者の組合せから得られる重みを評価値に加え、両者が一致しなかった場合にはその重みを評価値から減ずる。これにより、最終的にクラスフィールドの重みを加味した評価値が得られる。

【０１２３】また、類似事例抽出条件によっては、抽出される類似事例の数が多くなると実行時間が長くなる場合がある。このような場合、類似事例抽出条件から実行時間を類推し、実行時間の長さに応じた値を評価値から減ずることもできる。これにより、上述した評価値がほとんど変わらないいくつかの条件の中から、長い実行時間を要する条件を選択してしまうことが回避される。類推実行時間と実行時間を加味した評価値は、例えば、次式により与えられる。

$$（類推実行時間）＝（既知事例数）\times（フィールド数）＋\alpha\times（類似事例数）\quad（１１）$$

$$（評価値）＝（実行時間を加味する前の評価値）－\beta\times（類推実行時間）\quad（１２）$$

（１１）式において、類似事例数が類似事例抽出条件のみから定まらない場合は、条件付き類似事例抽出部３２から出力された類似事例集合の事例数の平均をとればよい。また、パラメータαは、類似事例抽出条件を基に決められ、パラメータβは、ユーザが実行時間をどれほど重要と認識するかを基に決められる。

【０１２４】次に、最適条件選択部３５は、条件評価部３４から出力された評価値に基づいて、離散化された類似事例抽出条件の中から最適な条件を選択する。例えば、類似事例抽出条件として類似事例数が指定され、横

軸を類似事例数とし縦軸を評価値として図４１に示すような評価値の分布が得られたものとする。

【０１２５】このとき、最適条件選択部３５は、評価値が最良となる条件を最適条件として選択するため、円で囲まれた最大値に対応する類似事例数が選択される。例えば、図４０に示した評価値の場合、最大値“３”に対応する“類似事例数＝３”が最適条件として選択される。

【０１２６】また、最適条件選択部３５は、評価値の移動平均を用いて最適な条件を選択することもできる。この場合、最適条件選択部３５は、条件ごとに適当な領域を定めて、その領域内に含まれる複数の条件の評価値について平均をとる。そして、得られた平均値を対応する条件の新たな評価値とし、その評価値が最大となる条件を最適条件として選択する。このような選択方法によれば、ノイズにより評価値が細かく振動しているような場合でも、それが滑らかに平均化されるため、より安定な条件を選択することができる。

【０１２７】例えば、図４１の評価値分布の場合、移動平均により図４２のような平均評価値が得られる。ここでは、基準となる条件の類似事例数ｋに対応する移動平均の領域を、ＭＡＸ（１，ｋ－２）以上ＭＩＮ（ｋ＿ｍａｘ，ｋ＋２）以下としている。ただし、ＭＡＸ（）、ＭＩＮ（）は、それぞれ（）内の数値の大きい方、小さい方をとるものとし、ｋ＿ｍａｘは類似事例数の最大値とする。

【０１２８】そして、ｋの値を１からｋ＿ｍａｘまでインクリメントしながら、この領域内の類似事例数に対する評価値の平均値を求めていくと、破線で示した平均評価値の分布が得られる。このとき、円で囲まれた平均評価値に対応する類似事例数が最適条件として選択される。

【０１２９】また、最適条件選択部３５は、評価値の近似関数を用いて最適な条件を選択することもできる。この場合、最適条件選択部３５は、すべての条件に渡る評価値の分布を適当な関数で近似し、その関数の値が最大となる条件を選択する。この近似関数としては、例えば、条件を記述するパラメータのｎ次多項式が用いられる。ｎは、１以上の整数であるが、実験的には、４、５等の値が適当であることが分かっている。

【０１３０】例えば、図４１の評価値分布の場合、近似関数により図４３のような評価値が得られる。ここでは、類似事例数ｋの２次多項式を近似関数ｆ（ｋ）として用い、次のような値が最小となるように各項の係数を定めた。

【０１３１】

【数６】

$$\sum_i (y_i - f(k_i))^2 \qquad (13)$$

【０１３２】ただし、k_i と y_i は、それぞれ、図４１の類似事例数と対応する評価値を表す。その結果、図４３に破線で示した近似関数が得られ、円で囲まれた関数値に対応する類似事例数が最適条件として選択される。

【０１３３】上述した各選択方法は、類似事例数だけでなく、類似度等の他の任意の類似事例抽出条件にも同様に適用される。類似事例数と類似度のように、異なる２つ以上の変数を含む類似事例抽出条件が与えられた場合は、変数ごとに独立に近似関数を求める等の方法により、図４３の選択方法を採用することが可能である。

【０１３４】こうして決定された最適な類似事例抽出条件は、図２に示したように、予測部１３に入力され、予測部１３は、入力された条件を用いて未知事例集合Ｕに関する予測を行う。このように、図２の構成によれば、既知事例集合Ａと未知事例集合Ｄを指定することで、自動的に最適な類似事例抽出条件が決定され、その条件を用いた予測が行われる。

【０１３５】ところで、予測部１３に含まれる図６の条件付き類似度計算部５１は、類似度計算において、既存のＣＣＦ法以外に他の任意の影響度計算方法を用いることができる。図３の類似事例抽出部２２についても同様である。以下では、既知事例集合のクラス値の分布の影響を受け、クラス値の分布の変化による重みへの影響がＣＣＦ法よりも大きくなるような重み付けを用いる計算方法について説明する。

【０１３６】まず、フィールドｊの値が領域ｖに含まれているときに、クラス値が領域ｃに含まれている条件付き確率をｐ（ｊ，ｖ，ｃ）とし、クラス値が領域ｃに含まれる確率をｐ（ｃ）とし、クラス値の数をＮ（ｃ）とする。例えば、２値のクラスフィールドについては、Ｎ（ｃ）＝２である。このとき、フィールドｊの重みｗ（ｊ，ｖ）は、（４）式の代わりに次式で与えられる。

【０１３７】

【数７】

$$q(j,v,c) = p(j,v,c)/p(c) \qquad (14)$$

$$w(j,v) = \frac{\sum_c \left| \frac{q(j,v,c)}{\sum_d q(j,v,d)} - \frac{1}{N(c)} \right|}{2 - \frac{2}{N(c)}} \qquad (15)$$

【０１３８】ここで、ｐ（ｊ，ｖ，ｃ）は、既知事例集合をフィールドｊの値により部分集合に分割した場合の領域ｖに対応する部分集合内のクラス値の分布に対応し、ｐ（ｃ）は全体のクラス値の分布に対応する。したがって、（１４）式のｑ（ｊ，ｖ，ｃ）は、既知事例集合の部分集合のクラス値の分布と全体のクラス値の分布の比を表している。

【０１３９】ｐ（ｊ，ｖ，ｃ）がｐ（ｃ）と等しいとき、ｑ（ｊ，ｖ，ｃ）＝１となり、（１５）式の分子は０となる。したがって、このとき、重みｗ（ｊ，ｖ）は最小値０をとる。また、特定のクラス値ｃについてのみ

ｐ（ｊ，ｖ，ｃ）＝１となり、ｃが他の値のときｐ（ｊ，ｖ，ｃ）＝０となるような場合は、（１５）式の分子は最大値２－２／Ｎ（ｃ）をとる。（１５）式の分母はこの最大値に一致しているため、このとき、重みｗ（ｊ，ｖ）は最大値１をとる。

【０１４０】言い換えれば、部分集合のクラス値の分布が全体のクラス値の分布に近いほど、対応するフィールドの影響度は小さく、部分集合のクラス値の分布が全体のクラス値の分布から遠いほど、対応するフィールドの影響度は大きくなる。この影響度計算方法による重みｗ（ｊ，ｖ）を（６）式のｗ（ｊ，ｖ（ｊ））として用いれば、既知事例集合のクラス値の分布を類似度計算に反映させることができ、既知事例集合のクラス値の分布が偏っているような場合でも、高い精度の予測が可能になる。

【０１４１】次に、図６の類似事例抽出部４１の動作について詳細に説明する。ここでは、類似事例数がｋであるという類似事例抽出条件が与えられたものとする。条件付き類似度計算部５１は、類似度計算の過程で、上述した（６）式の平方根の中の総和を求めるとき、フィールド１に対応する項から順番に加算していく。

【０１４２】この加算において、総和は単調に増加し、それに伴って類似度Ｓは単調に減少する。したがって、既知事例を新しく類似事例集合に加えるための類似度条件が“Ｓ≧Ｓ１”であるとすると、Ｓ＜Ｓ１となった時点で、それ以上加算を続けてもその既知事例は類似事例になり得ないことが分かる。

【０１４３】そこで、条件付き類似度計算部５１は、一定間隔でＳ≧Ｓ１であるかどうかをチェックし、この条件が満たされないときには類似度計算を中止して、次の既知事例の類似度計算を開始する。そして、類似度条件を満たす既知事例を新たな類似事例として出力する。

【０１４４】類似事例集合更新部５２は、類似事例集合記憶部５３から現在の類似事例集合を取り出し、類似事例抽出条件に従って、条件付き類似度計算部５１から出力された類似事例を類似事例集合に加える。このとき、新しい類似事例集合の事例数がｋ以下であれば、その類似事例集合を更新結果として出力し、新しい類似事例集合の事例数がｋ＋１であれば、類似度が最も小さい事例を削除し、得られた類似事例集合を更新結果として出力する。

【０１４５】類似事例集合記憶部５３は、類似事例集合更新部５２から出力された類似事例集合を記憶する。ただし、初期状態においては、類似事例集合は空集合である。類似度条件計算部５４は、類似事例集合記憶部５３内の類似事例集合の事例数がｋであるとき、“Ｓ≧Ｓ１”という類似度条件を出力する。Ｓ１としては、例えば、類似事例集合内の事例の類似度の最小値が用いられる。

【０１４６】また、類似事例集合の事例数がｋ未満であるときは、“条件なし”を出力する。この場合、条件付き類似度計算部５１は、類似度計算を途中で打ち切ることなく、どんな類似度の事例であっても類似事例として出力する。

【０１４７】このように、類似事例抽出条件と既に得られている類似事例の類似度に応じて類似度計算を中止するための条件を決定し、その条件に従って計算を中止することで、類似事例抽出の効率が向上する。

【０１４８】次に、図２の予測装置１１の主要部により行われる処理の例について、フローチャートを参照しながらより詳細に説明する。

【０１４９】図４４は、図８の事例削除部６２の処理のフローチャートである。事例削除部６２は、まず、未知事例入力用事例集合Ｃの事例の番号を表す制御変数Ｉを１とおき（ステップＳ１）、Ｉを未知事例入力用事例集合Ｃの事例数｜Ｃ｜と比較する（ステップＳ２）。ここでは、未知事例入力用事例集合Ｃは既知事例集合Ａに一致している。

【０１５０】Ｉ≦｜Ｃ｜であれば、集合ＣのＩ番目の事例Ｃ［Ｉ］を既知事例集合Ａから削除して、事例Ｃ［Ｉ］に対応する既知事例入力用事例集合Ｂ［Ｉ］を生成し、Ｉ＝Ｉ＋１とおいて（ステップＳ３）、ステップＳ２以降の処理を繰り返す。そして、ステップＳ２においてＩ＞｜Ｃ｜となると、得られた集合Ｂ［Ｉ］を既知事例入力用事例集合Ｂとして出力し、処理を終了する。

【０１５１】次に、図４５は、図３の類似事例削除部２３の処理のフローチャートである。類似事例削除部２３は、まず、制御変数Ｉを１とおき（ステップＳ１１）、Ｉを未知事例入力用事例集合Ｃの事例数｜Ｃ｜と比較する（ステップＳ１２）。

【０１５２】Ｉ≦｜Ｃ｜であれば、集合ＣのＩ番目の事例Ｃ［Ｉ］を、Ｃ［Ｉ］に対応する類似事例集合Ｎ［Ｉ］から削除して、修正された類似事例集合Ｍ［Ｉ］を生成し、Ｉ＝Ｉ＋１とおいて（ステップＳ１３）、ステップＳ１２以降の処理を繰り返す。そして、ステップＳ１２においてＩ＞｜Ｃ｜となると、得られた集合Ｍ［Ｉ］を出力して、処理を終了する。

【０１５３】次に、図４６は、図４の条件付き類似事例抽出部３２の処理のフローチャートである。条件付き類似事例抽出部３２は、まず、制御変数Ｉを１とおき（ステップＳ２１）、Ｉを未知事例入力用事例集合Ｃの事例数｜Ｃ｜と比較する（ステップＳ２２）。

【０１５４】Ｉ≦｜Ｃ｜であれば、離散化された条件の番号を表す制御変数Ｘを１とおき（ステップＳ２３）、Ｘを離散化された条件の数Ｎと比較する（ステップＳ２４）。Ｘ≦Ｎであれば、事例Ｃ［Ｉ］に対応する修正された類似事例集合Ｍ［Ｉ］の事例の番号を表す制御変数Ｙを１とおき、Ｃ［Ｉ］およびＸ番目の条件に対応する条件ごとの類似事例集合Ｐ［Ｉ］［Ｘ］を空集合φとおいて（ステップＳ２５）、Ｙを集合Ｍ［Ｉ］の事例数｜

M[I]|と比較する（ステップS26）。
【0155】Y≦|M[I]|であれば、集合M[I]のY番目の事例M[I][Y]がX番目の条件を満たすかどうかをチェックする（ステップS27）。事例M[I][Y]がX番目の条件を満たせば、その事例を集合P[I][X]に加え（ステップS28）、Y=Y+1とおいて（ステップS29）、ステップS26以降の処理を繰り返す。また、事例M[I][Y]がX番目の条件を満たさなければ、その事例を集合P[I][X]に加えずに、ステップS29以降の処理を繰り返す。
【0156】次に、ステップS26においてY>|M[I]|となると、X=X+1とおいて（ステップS30）、ステップS24以降の処理を繰り返す。次に、ステップS24においてX>Nとなると、I=I+1とおいて（ステップS31）、ステップS22以降の処理を繰り返す。そして、ステップS22においてI>|C|となると、得られた類似事例集合P[I][X]を出力して、処理を終了する。
【0157】次に、図47は、図4の条件評価部34の処理のフローチャートである。ここでは、未知事例入力用事例集合Cの事例のクラス値がカテゴリ値である場合を考え、真のクラス値と一致した予測値の数を評価値として用いている。
【0158】条件評価部34は、まず、制御変数Xを1とおき（ステップS41）、Xを条件の数Nと比較する（ステップS42）。X≦Nであれば、制御変数Iを1とおき、X番目の条件の評価値E[X]を0とおいて（ステップS43）、Iを未知事例入力用事例集合Cの事例数|C|と比較する（ステップS44）。
【0159】I≦|C|であれば、X番目の条件における事例C[I]の予測値R[I][X]をC[I]のクラス値と比較する（ステップS45）。R[I][X]がC[I]のクラス値と一致すれば、E[X]に1を加算し（ステップS46）、I=I+1とおいて（ステップS47）、ステップS44以降の処理を繰り返す。また、R[I][X]がC[I]のクラス値と一致しなければ、E[X]を更新せずに、ステップS47以降の処理を繰り返す。
【0160】次に、ステップS44においてI>|C|となると、X=X+1とおいて（ステップS48）、ステップS42以降の処理を繰り返す。そして、ステップS42においてX>Nとなると、得られた評価値E[X]を出力して、処理を終了する。
【0161】次に、図48は、図4の最適条件選択部35の処理のフローチャートである。ここでは、図42に示した移動平均に基づく選択方法を用いている。最適条件選択部35は、まず、制御変数Xを1とおき、最適条件の番号を表す制御変数MAXを1とおいて（ステップS51）、Xを条件の数Nと比較する（ステップS52）。
【0162】X≦Nであれば、X番目の条件を基準とする所定領域に含まれる条件の集合をS[X]として、集合S[X]内の条件の番号を表す制御変数Zを1とおき、X番目の条件の平均評価値F[X]を0とおく（ステップS53）。そして、Zを集合S[X]の条件数|S[X]|と比較する（ステップS54）。
【0163】Z≦|S[X]|であれば、S[X]のZ番目の条件S[X][Z]の評価値E[S[X][Z]]をF[X]に加算し、Z=Z+1とおいて（ステップS55）、ステップS54以降の処理を繰り返す。これにより、S[X]に含まれる条件の評価値の総和がF[X]に格納される。
【0164】次に、Z>|S[X]|となると、F[X]を|S[X]|で割って平均評価値を求め、それをF[X]に格納して（ステップS56）、F[X]とF[MAX]を比較する（ステップS57）。F[X]>F[MAX]であれば、MAX=Xとおき（ステップS58）、X=X+1とおいて（ステップS59）、ステップS52以降の処理を繰り返す。また、F[X]≦F[MAX]であれば、MAXを更新せずに、ステップS59以降の処理を繰り返す。
【0165】そして、ステップS52においてX>Nとなると、平均評価値の最大値はF[MAX]であるので、対応するMAX番目の条件を最適条件として出力して（ステップS60）、処理を終了する。
【0166】次に、図49は、図6の条件付き類似度計算部51の処理のフローチャートである。ここでは、上述した（6）式に従って既知事例と未知事例の類似度Sを計算しており、類似度条件として“S≧S1”を用いている。
【0167】条件付き類似度計算部51は、まず、（6）式の平方根の中の総和を表す変数Dを0とおき、既知事例のフィールドの番号を表す制御変数Jを1とおいて（ステップS61）、$w(J, v(J))d(J)^2$をDに加算する（ステップS62）。このとき、類似度条件“$S=1/(D)^{1/2} \geqq S1$”は“$D \leqq 1/S1^2$”と書き換えられるので、Dを$1/S1^2$と比較する（ステップS63）。
【0168】$D \leqq 1/S1^2$であれば、Jをフィールドの数Nfと比較する（ステップS64）。J<Nfであれば、J=J+1とおいて（ステップS65）、ステップS62以降の処理を繰り返す。そして、ステップS64においてJ=Nfとなると、得られたDの平方根の逆数を類似度Sとおいて（ステップS66）、処理を終了する。
【0169】また、ステップS63においてD>$1/S1^2$となると、その既知事例は類似事例になり得ないものと判断し（ステップS67）、類似度計算を中止して、処理を終了する。
【0170】以上説明した図2の予測装置11は、任意

のデータ分類処理に適用することができる。図５０は、予測装置１１を含むデータ分類装置の構成図である。図５０のデータ分類装置は、予測装置１１、既知事例データベース８１、未知事例データベース８２、入力装置８３、分類装置８４、および出力装置８５を備える。

【０１７１】既知事例データベース８１と未知事例データベース８２は、それぞれ、既知事例集合と未知事例集合を格納し、入力装置８３は、既知事例データベース８１と未知事例データベース８２から予測装置１１に事例集合を入力する。予測装置１１は、既知事例集合を用いて、各未知事例のクラスフィールドを予測し、予測結果を出力する。分類装置８４は、クラスフィールドの予測値に従って未知事例を分類し、出力装置８５は、分類結果をディスプレイ画面等に出力する。

【０１７２】例えば、図１２に示したようなフィールド構成を持つ未知事例の場合は、クラスフィールド“応答”の予測値が“あり”と“なし”のいずれであるかに応じて、２つのグループに分類され、予測値“あり”のグループがダイレクトメールの送り先として出力される。このとき、出力装置８５は、単に分類結果を出力するだけでなく、通信ネットワーク８６に自動的に接続して、指定された送り先に適当なメッセージを含む電子メールを送信することもできる。

【０１７３】このようなデータ分類装置によれば、多数の未知事例の中からダイレクトメールやアンケートの対象者、金融機関による貸し出し先等を決定したり、保険の契約者を未知事例として分類したりすることができる。

【０１７４】また、装置やネットワークの構成要素を未知事例とし、故障の有無をクラスフィールドとして分類を行うことで、故障箇所を推定することもできる。この場合、出力装置８５は、例えば、故障箇所と推定された構成要素に適当な制御信号や制御メッセージを送ってその要素を制御し、復旧処理を行う。

【０１７５】図５０のデータ分類装置は、図５１に示すような情報処理装置（コンピュータ）を用いて構成することができる。図５１の情報処理装置は、ＣＰＵ（中央処理装置）９１、メモリ９２、入力装置９３、出力装置９４、外部記憶装置９５、媒体駆動装置９６、およびネットワーク接続装置９７を備え、それらはバス９８により互いに接続されている。

【０１７６】メモリ９２は、例えば、ＲＯＭ（read only memory）、ＲＡＭ（random access memory）等を含み、処理に用いられるプログラムとデータを格納する。ＣＰＵ９１は、メモリ９２を利用してプログラムを実行することにより、必要な処理を行う。

【０１７７】図２の類似事例抽出条件決定部１２および予測部１３、図３の入力用事例生成部２１、類似事例抽出部２２、類似事例削除部２３、最適条件決定部２４、条件出力部２５、最大条件計算部２６、および最大条件修正部２７、図４の条件離散化部３１、条件付き類似事例抽出部３２、予測結果生成部３３、条件評価部３４、および最適条件選択部３５、図５の類似事例抽出部４１および予測結果生成部４２、図６の条件付き類似度計算部５１、類似事例集合更新部５２、類似事例集合記憶部５３、および類似度条件計算部５４、図５０の分類装置８４等は、メモリ９２の特定のプログラムコードセグメントに格納されたソフトウェアコンポーネントに対応する。

【０１７８】入力装置９３は、例えば、キーボード、ポインティングデバイス、タッチパネル等であり、ユーザからの指示や情報の入力に用いられる。出力装置９４は、例えば、モニタディスプレイ、プリンタ等であり、ユーザへの問い合わせや処理結果等の出力に用いられる。

【０１７９】外部記憶装置９５は、例えば、磁気ディスク装置、光ディスク装置、光磁気ディスク（magneto-optical disk）装置等である。この外部記憶装置９５に、上述のプログラムとデータを保存しておき、必要に応じて、それらをメモリ９２にロードして使用することもできる。

【０１８０】媒体駆動装置９６は、可搬記録媒体９９を駆動し、その記録内容にアクセスする。可搬記録媒体９９としては、メモリカード、フロッピーディスク、ＣＤ－ＲＯＭ（compact disk read only memory）、光ディスク、光磁気ディスク等、任意のコンピュータ読み取り可能な記録媒体が用いられる。この可搬記録媒体９９に上述のプログラムとデータを格納しておき、必要に応じて、それらをメモリ９２にロードして使用することもできる。

【０１８１】ネットワーク接続装置９７は、ＬＡＮ（local area network）等の任意のネットワーク（回線）を介して外部の装置と通信し、通信に伴うデータ変換を行う。また、必要に応じて、上述のプログラムとデータを外部の装置から受け取り、それらをメモリ９２にロードして使用することもできる。

【０１８２】例えば、図５０の既知事例データベース８１、未知事例データベース８２、および入力装置８３は外部記憶装置９５に対応し、図５０の出力装置８５は、出力装置９４およびネットワーク接続装置９７に対応する。

【０１８３】図５２は、図５１の情報処理装置にプログラムとデータを供給することのできるコンピュータ読み取り可能な記録媒体を示している。可搬記録媒体９９や外部のデータベース１００に保存されたプログラムとデータは、メモリ９２にロードされる。そして、ＣＰＵ９１は、そのデータを用いてそのプログラムを実行し、必要な処理を行う。

【０１８４】

【発明の効果】本発明によれば、類似事例に基づく予測

において、既知事例集合に対して前処理を行うことなく、予測を高速化することができる。また、類似事例の抽出を何度も繰り返すことなく、高速かつ自動的に良好な類似事例抽出条件が得られる。さらに、既知事例集合のクラス値の分布が偏っているような場合でも、高い精度の予測が可能になる。

【0185】このように、本発明によれば、高速かつ高精度な予測が実現され、多様なデータ分類処理に寄与するところが大きい。

【図面の簡単な説明】

【図1】本発明の予測装置の原理図である。
【図2】予測装置の構成図である。
【図3】類似事例抽出条件決定部の構成図である。
【図4】最適条件決定部の構成図である。
【図5】予測部の構成図である。
【図6】類似事例抽出部の構成図である。
【図7】第1の入力用事例生成部を示す図である。
【図8】第2の入力用事例生成部を示す図である。
【図9】第3の入力用事例生成部を示す図である。
【図10】第4の入力用事例生成部を示す図である。
【図11】第5の入力用事例生成部を示す図である。
【図12】既知事例集合を示す図である。
【図13】未知事例入力用事例集合を示す図である。
【図14】第1の類似事例集合を示す図である。
【図15】第2の類似事例集合を示す図である。
【図16】第3の類似事例集合を示す図である。
【図17】第4の類似事例集合を示す図である。
【図18】第5の類似事例集合を示す図である。
【図19】第6の類似事例集合を示す図である。
【図20】第7の類似事例集合を示す図である。
【図21】第8の類似事例集合を示す図である。
【図22】第9の類似事例集合を示す図である。
【図23】第10の類似事例集合を示す図である。
【図24】第11の類似事例集合を示す図である。
【図25】第12の類似事例集合を示す図である。
【図26】第13の類似事例集合を示す図である。
【図27】第14の類似事例集合を示す図である。
【図28】第15の類似事例集合を示す図である。
【図29】第16の類似事例集合を示す図である。
【図30】第17の類似事例集合を示す図である。
【図31】第18の類似事例集合を示す図である。
【図32】第19の類似事例集合を示す図である。
【図33】第20の類似事例集合を示す図である。
【図34】第21の類似事例集合を示す図である。
【図35】第22の類似事例集合を示す図である。
【図36】第23の類似事例集合を示す図である。
【図37】第24の類似事例集合を示す図である。
【図38】予測結果を示す図である。
【図39】予測結果生成部の構成図である。
【図40】評価値を示す図である。
【図41】第1の最適条件を示す図である。
【図42】第2の最適条件を示す図である。
【図43】第3の最適条件を示す図である。
【図44】事例削除部の処理のフローチャートである。
【図45】類似事例削除部の処理のフローチャートである。
【図46】条件付き類似事例抽出部の処理のフローチャートである。
【図47】条件評価部の処理のフローチャートである。
【図48】最適条件選択部の処理のフローチャートである。
【図49】条件付き類似度計算部の処理のフローチャートである。
【図50】データ分類装置の構成図である。
【図51】情報処理装置の構成図である。
【図52】記録媒体を示す図である。

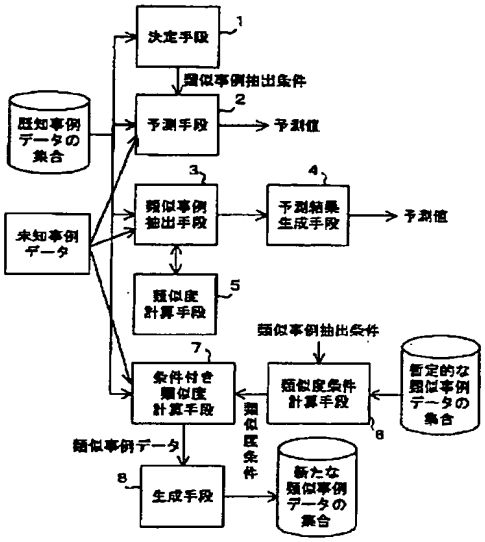
【符号の説明】

1　決定手段
2　予測手段
3　類似事例抽出手段
4　予測結果生成手段
5　類似度計算手段
6　類似度条件計算手段
7　条件付き類似度計算手段
8　生成手段
11　予測装置
12　類似事例抽出条件決定部
13　予測部
21　入力用事例生成部
22、41　類似事例抽出部
23　類似事例削除部
24　最適条件決定部
25　条件出力部
26　最大条件計算部
27　最大条件修正部
31　条件離散化部
32　条件付き類似事例抽出部
33、42　予測結果生成部
34　条件評価部
35　最適条件選択部
51　条件付き類似度計算部
52　類似事例集合更新部
53　類似事例集合記憶部
54　類似度条件計算部
61　分割部
62　事例削除部
63　サンプリング部
71　一次予測生成部
72　予測修正部
81、82、100　データベース

83、93　入力装置
84　分類装置
85、94　出力装置
86　ネットワーク
91　CPU
95　外部記憶装置
96　媒体駆動装置
97　ネットワーク接続装置
98　バス
99　可搬記録媒体
//

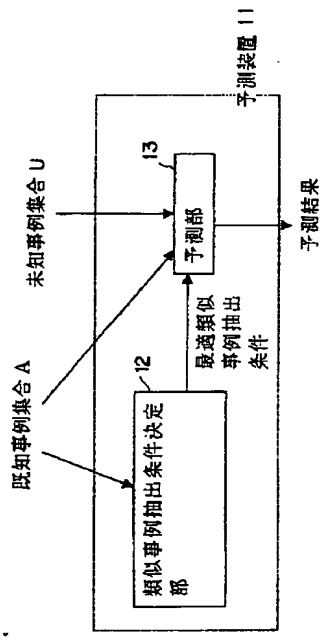
【図1】

本　発　明　の　原　理　図



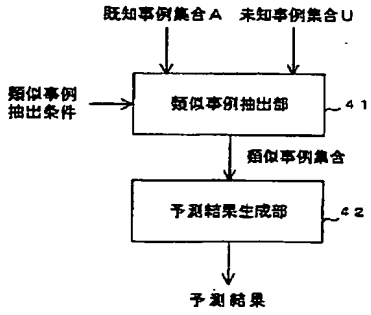
【図2】

予測装置の構成図



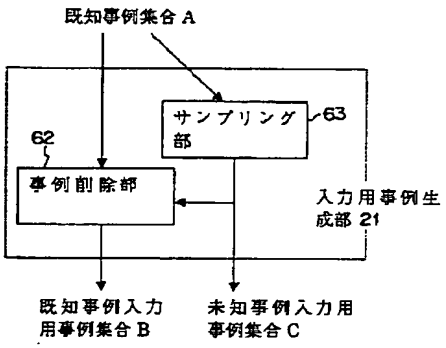
【図5】

予　測　部　の　構　成　図



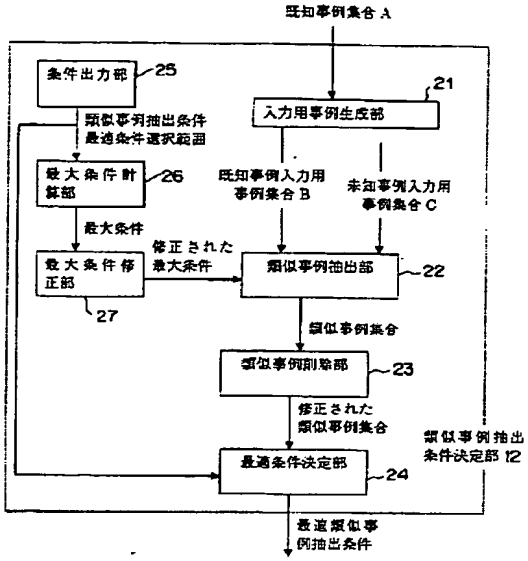
【図10】

第4の入力用事例生成部を示す図



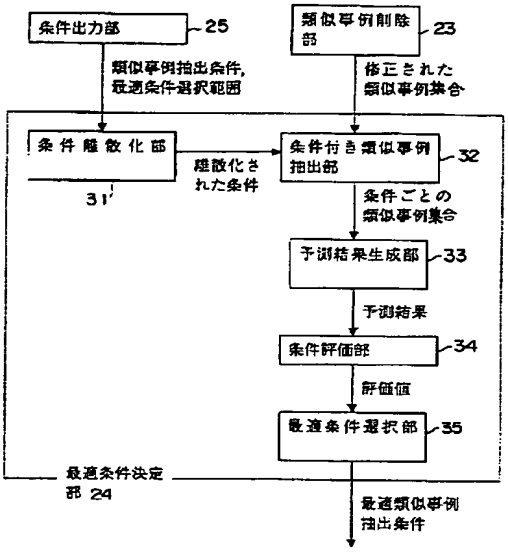
【図３】

類似事例抽出条件決定部の構成図



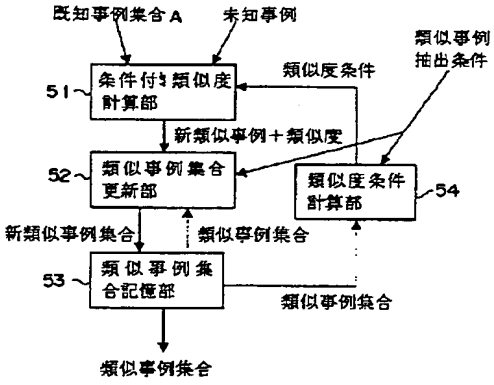
【図４】

最適条件決定部の構成図



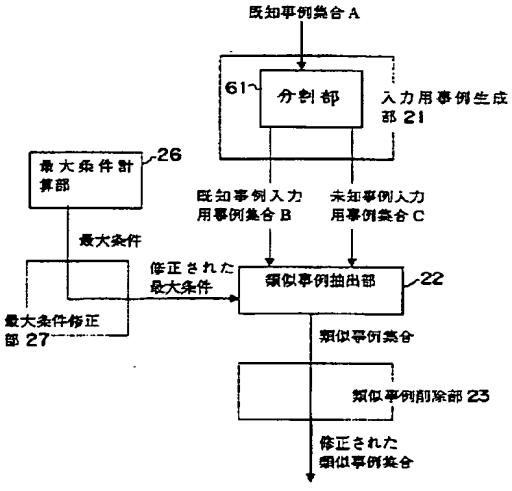
【図６】

類似事例抽出部の構成図



【図７】

第１の入力用事例生成部を示す図



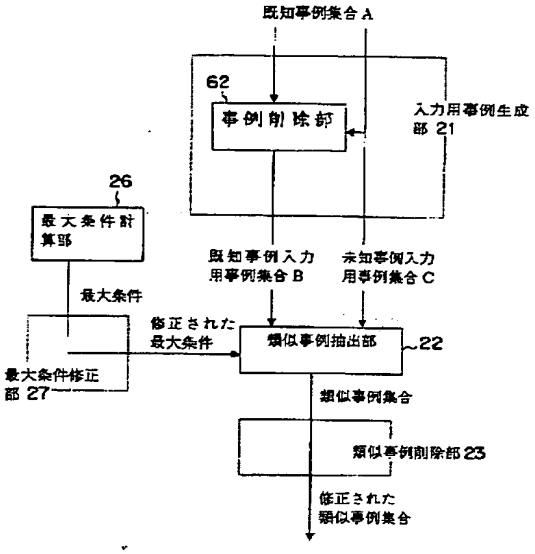
【図１３】

未知事例入力用事例集合を示す図

氏名	年齢	性別	職業	結婚	応答
A	30	男	公務員	既婚	あり
C	40	女	無職	既婚	なし
E	22	男	学生	未婚	あり
G	34	男	会社員	未婚	なし

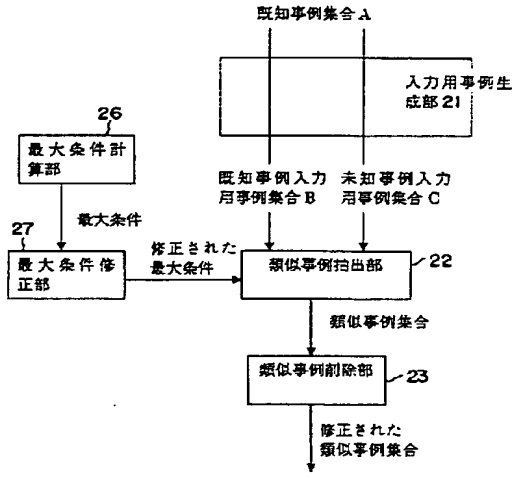
【図8】

第2の入力用事例生成部を示す図



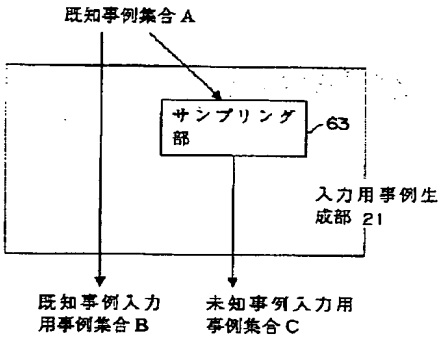
【図9】

第3の入力用事例生成部を示す図



【図11】

第5の入力用事例生成部を示す図



【図12】

既知事例集合を示す図

氏名	年齢	性別	職業	結婚	応答
A	30	男	公務員	既婚	あり
B	25	男	会社員	未婚	あり
C	40	女	無職	既婚	なし
D	50	男	農業	既婚	なし
E	22	男	学生	未婚	あり
F	20	女	学生	未婚	なし
G	34	男	会社員	未婚	なし

【図14】

第1の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	****
G	34	男	会社員	未婚	なし	6
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	2
D	50	男	農業	既婚	なし	1.5

【図26】

第13の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	6
B	25	男	会社員	未婚	あり	4

【図15】

第2の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
C	40	女	無職	既婚	なし	***
G	34	男	会社員	未婚	なし	5
A	30	男	公務員	既婚	あり	3
D	50	男	農業	既婚	なし	2.5
F	20	女	学生	未婚	なし	2
B	25	男	会社員	未婚	あり	1

【図16】

第3の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
E	22	男	学生	未婚	あり	***
B	25	男	会社員	未婚	あり	6
A	30	男	公務員	既婚	あり	3
F	20	女	学生	未婚	なし	2.5
G	34	男	会社員	未婚	なし	1.5
C	40	女	無職	既婚	なし	1

【図17】

第4の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	***
A	30	男	公務員	既婚	あり	5
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	1.5
D	50	男	農業	既婚	なし	1

【図18】

第5の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	6
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	2
D	50	男	農業	既婚	なし	1.5

【図19】

第6の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	5
A	30	男	公務員	既婚	あり	3
D	50	男	農業	既婚	なし	2.5
F	20	女	学生	未婚	なし	2
B	25	男	会社員	未婚	あり	1

【図20】

第7の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
B	25	男	会社員	未婚	あり	6
A	30	男	公務員	既婚	あり	3
F	20	女	学生	未婚	なし	2.5
G	34	男	会社員	未婚	なし	1.5
C	40	女	無職	既婚	なし	1

【図21】

第8の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	5
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	1.5
D	50	男	農業	既婚	なし	1

【図22】

第9の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	6

【図23】

第10の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	5

【図24】

第11の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
B	25	男	会社員	未婚	あり	6

【図25】

第12の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	5

【図27】

第14の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	5
A	30	男	公務員	既婚	あり	3

【図28】

第15の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
B	25	男	会社員	未婚	あり	6
A	30	男	公務員	既婚	あり	3

【図29】

第16の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	5
B	25	男	会社員	未婚	あり	4

【図30】

第17の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	6
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3

【図31】

第18の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	5
A	30	男	公務員	既婚	あり	3
D	50	男	農業	既婚	なし	2.5

【図32】

第19の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
B	25	男	会社員	未婚	あり	6
A	30	男	公務員	既婚	あり	3
F	20	女	学生	未婚	なし	2.5

【図33】

第20の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	5
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3

【図34】

第21の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	6
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	2

【図35】

第22の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
G	34	男	会社員	未婚	なし	5
A	30	男	公務員	既婚	あり	3
D	50	男	農業	既婚	なし	2.5
F	20	女	学生	未婚	なし	2

【図40】

評価値を示す図

類似事例数	1	2	3	4	5
評価値	2	2	3	2	2

【図36】

第23の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
B	25	男	会社員	未婚	あり	8
A	30	男	公務員	既婚	あり	3
F	20	女	学生	未婚	なし	2.5
G	34	男	会社員	未婚	なし	1.5

【図37】

第24の類似事例集合を示す図

氏名	年齢	性別	職業	結婚	応答	類似度
A	30	男	公務員	既婚	あり	5
B	25	男	会社員	未婚	あり	4
E	22	男	学生	未婚	あり	3
C	40	女	無職	既婚	なし	1.5

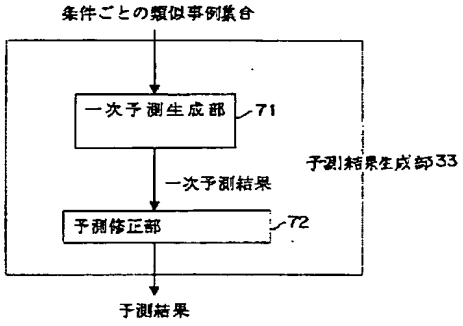
【図38】

予測結果を示す図

類似事例数	1	2	3	4	5
A	なし	なし	あり	なし	なし
C	なし	なし	なし	なし	なし
E	あり	あり	あり	あり	あり
G	あり	あり	あり	あり	あり

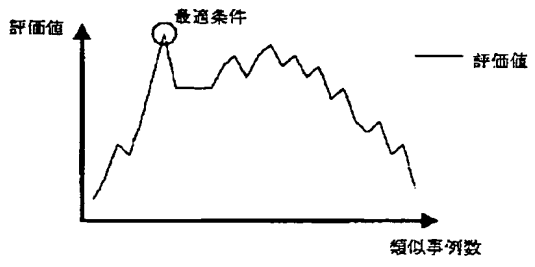
【図39】

予測結果生成部の構成部



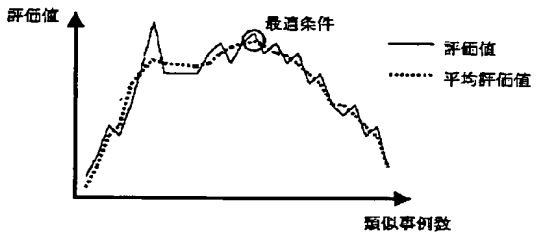
【図41】

第1の最適条件を示す図



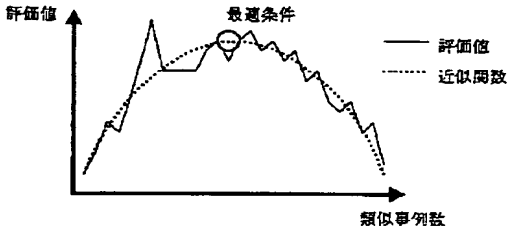
【図42】

第2の最適条件を示す図



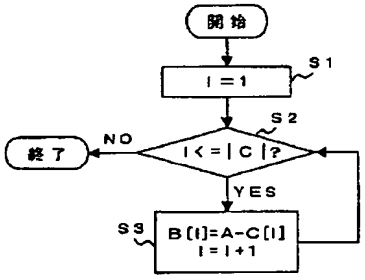
【図43】

第3の最適条件を示す図



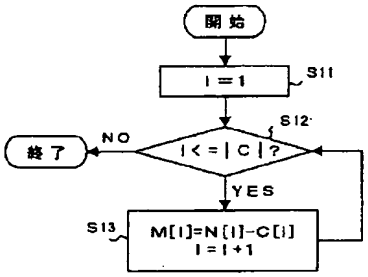
【図44】

事例削除部の処理のフローチャート



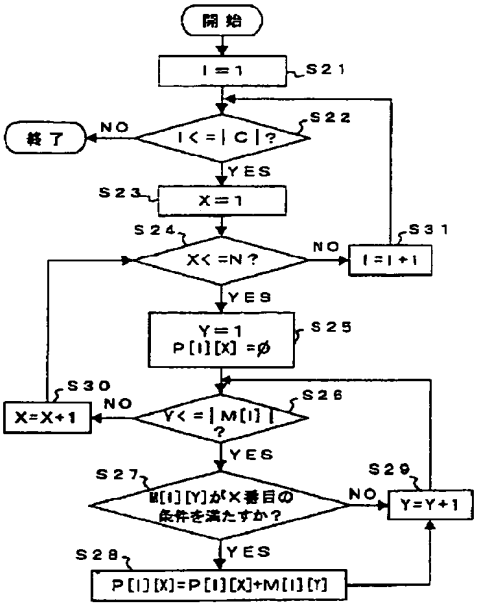
【図45】

類似事例削除部の処理のフローチャート



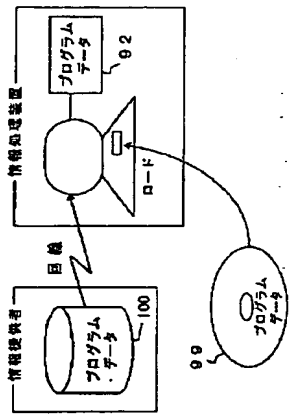
【図46】

条件付き類似事例抽出部の処理のフローチャート



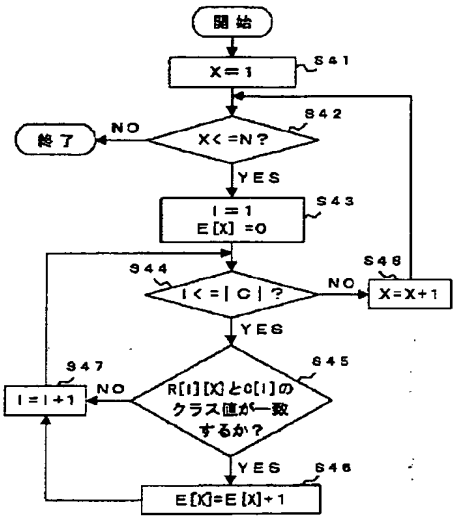
【図52】

記録媒体を示す図



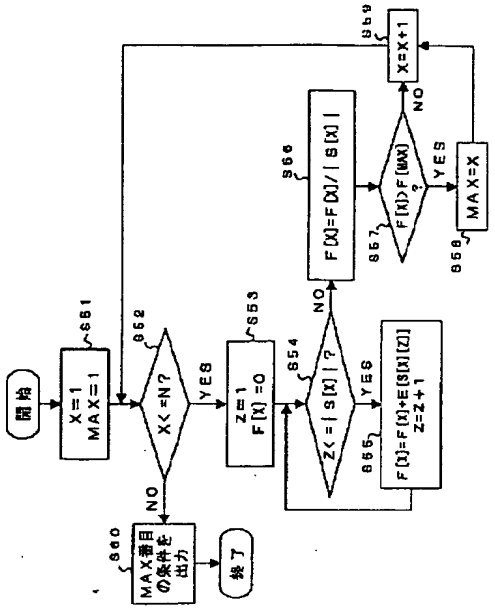
【図47】

条件評価部の処理のフローチャート



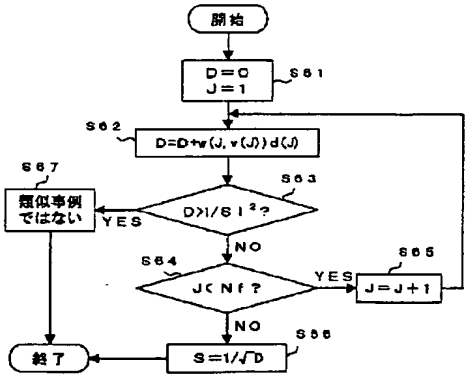
【図48】

最適条件選択部の処理のフローチャート



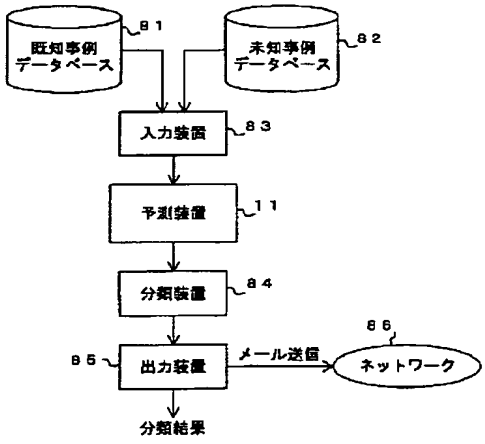
【図49】

条件付き類似度計算部の処理のフローチャート



【図50】

データ分類装置の構成図



【図51】

情報処理装置の構成図